KU LEUVEN

FACULTY OF ECONOMICS AND BUSINESS

Performance measurement of a rotavirus vaccine supply chain design by the integration of production capacity into the guaranteed service approach

Lemmens S, Decouttere C, Vandaele N, Bernuzzi M, Reichman A.

KBI_1621

Performance measurement of a rotavirus vaccine supply chain design by the integration of production capacity into the guaranteed service approach

Stef Lemmens*[1], Catherine J. Decouttere[1], Nico J. Vandaele[1], Mauro Bernuzzi[2], and Amir Reichman[2]

[1]KULeuven, Research Center for Operations Management, Naamsestraat 69, 3000 Leuven, Belgium
[2]GlaxoSmithKline Vaccines, Avenue Fleming 20, 1300 Waver, Belgium

Abstract

Previous research has integrated multi-echelon inventory management into the design of a responsive supply chain by the use of the guaranteed service approach. We build further upon this work by integrating the production capacity and product flow to minimize the supply chain's inventories. The production capacity is modeled with a queuing network to handle the variability of the batch production processes as well as the demand variability. We test and validate our model with adapted instances from literature and apply it to the rotavirus vaccine supply chain. This vaccine supply chain is seen as complex on the manufacturing side as well as on the distribution side. For our industrial application we show how this work is embedded in a scenario approach and the contribution of our model to evaluate a single scenario according to multiple performance indicators. For this paper, our scenarios consist of different lead time reduction programmes and varying demand levels. We demonstrate how to extract the best performing scenario.

1 Introduction

Emerging pressure from global competitors and stringent agreements with the final customer on the delivery date require a responsive end-to-end supply chain. We define such a responsive supply chain as a supply chain that is able to meet a fluctuating demand with variable, but relatively short lead times. Ideally, this means that a responsive supply chain is still able to satisfy the customer's demand in case of undesirable variability of the process durations (e.g. unexpected maintenance, material issues or quality problems) as well as variability stemming from externalities (e.g. tenders or natural disasters leading to sudden demand changes). Such undesirable variability of the process durations may induce a lead time distribution which can take a wide range of values and might expose a fat tail. However, supply chain responsiveness cannot only be improved by reducing the lead times of the different processes and their variability, but also by trading off operational buffers ([Vandaele and De Boeck, 2003]): such buffers include the position and the volume of the strategic stocks and the load of the installed production capacity in the supply chain.

A literature review on supply chain network design models by [Lemmens et al., 2016] shows that the majority of the literature imposes an economical performance criterion instead of a responsiveness criterion in the research field on supply chain design whilst literature confirms the importance of a lead time driven supply chain metric. Recent work of [de Treville et al., 2014] shows for three industrial supply chains (GSK Vaccines, Nissan Europe and Nestlé Switzerland) that managers underestimate the benefits of cutting lead times. Our industrial application is in line with Suri's work on Quick Response Manufacturing ([Suri, 1998]): this work describes a company wide approach to reduce lead times and emphasizes the industrial relevance.

The challenge and importance of the design of a responsive vaccine supply chain is emphasized by [Shah, 2004, Shah, 2005]. The vaccine industry is characterized by complex manufacturing processes and stringent regulatory processes which tend to be slow. The vaccine supply chain involves primary

*Corresponding author
E-mail address: stef.lemmens@kuleuven.be

processes (cultivation of the antigen), secondary processes (formulation, filling and packaging) and a meticulous process of continuous quality control and quality assurance. All this leads to an overall supply chain lead time of more than 300 days ([VaccinesEurope, 2015], [IFPMA, 2016]).

GSK Vaccines is located in Wavre (Belgium) and has 17 production facilities which manufactured and distributed more than 690 million doses across 170 countries in 2015 ([GSK, 2016b]). GSK Vaccines is partnering with Gavi, The Vaccine Alliance, and Save the Children and one of the company's deep commitments is to contribute to the accessibility of its vaccines through equitable or tiered pricing ([GSK, 2014]). Recently, at the meeting of the UN High Level Panel on Access to Medicines, GSK CEO Sir Andrew Witty set out a series of steps designed to help bring innovative GSK medicines to more people living in the world's poorest countries ([GSK, 2016a]). For the 20 leading pharmaceutical companies the Access to Medicines Index measures the performance of providing vaccines, medicines and healthcare in developing countries ([ATM, 2016]) and GSK has been placed multiple times at the top of this index ([GSK, 2014]).

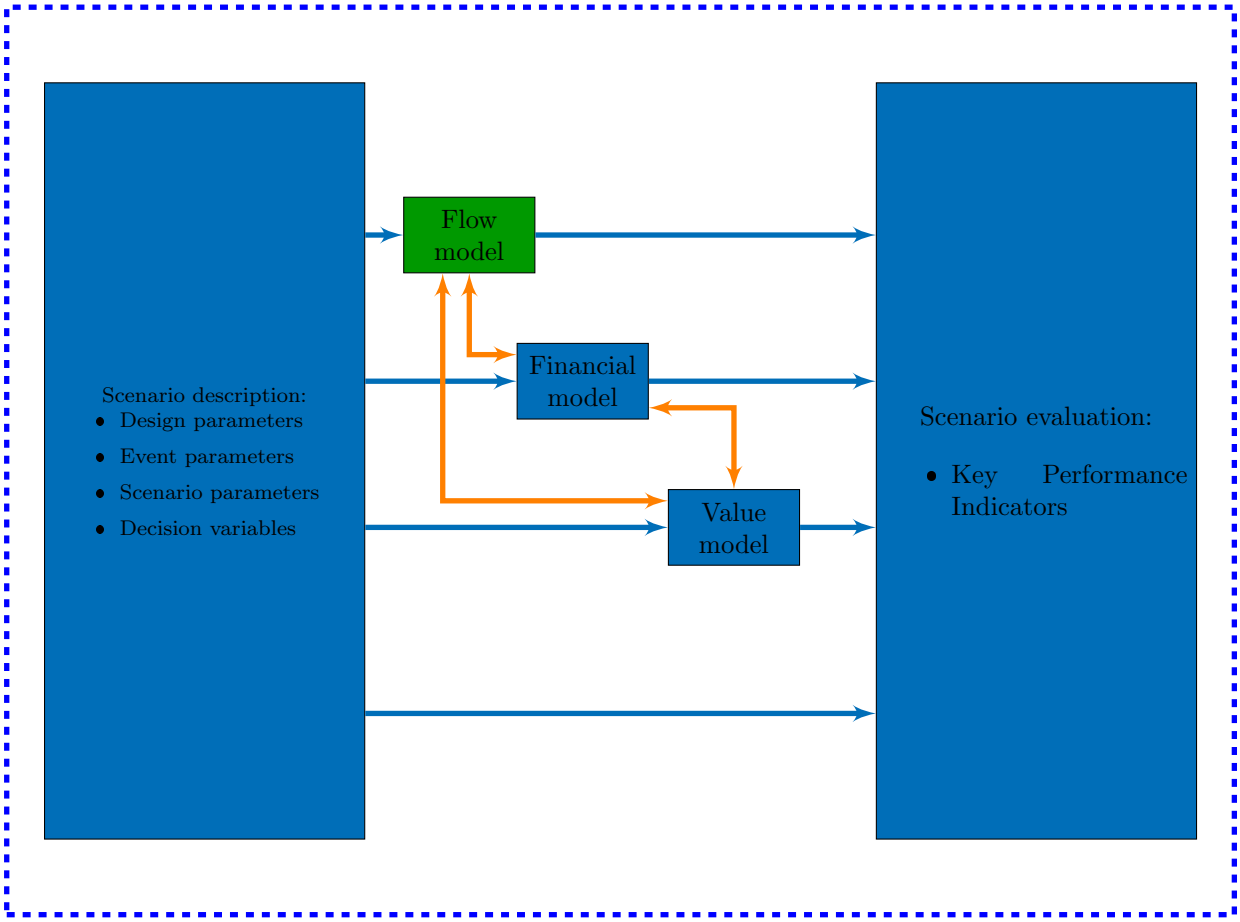
In general, the global demand for vaccines exceeds its supply and large volumes are negotiated by tender contracts. Furthermore vaccine supply chains face long and variable lead times and the regulatory affairs and vaccines' perishability limit potential stock building ([Lemmens et al., 2016]). For GSK Vaccines, [de Treville et al., 2014] explains that these long and variable lead times also resulted from a tactic to move production between factories in search of the highest capacity utilization to obtain cost savings.

We demonstrate our approach for the rotavirus vaccine manufacturing supply chain of GSK Vaccines. We selected this vaccine on the basis of the availability of both capacity and supply related data and expert knowledge from GSK Vaccines ([Decouttere et al., 2015]). The rotavirus vaccine has been introduced in the national immunization programme of many developing countries and allowed us to conduct an extended stakeholder analysis ([Decouttere et al., 2015]). In addition, the production lines for this vaccine are dedicated which allows us to consider this work as a pilot study before elaborating on more complex manufacturing supply chains. These complex manufacturing supply chains may include (1) shared capacity between different vaccines or (2) combination vaccines against multiple diseases such as the Diphtheria-Tetanus-Pertussis (DTP) vaccine ([WHO, 2016]).

For the rotavirus vaccine, the available production time of each manufacturing line is obtained by determining the quarterly available time (based on e.g. the number of working shifts per day and the number of holidays) and subsequently deducting the time allocated to maintenance, setup, and validation of the line. Given the remaining available time and the nominal speed of the production lines, it is identified whether or not the manufacturing supply chain is able to produce the demand volume. At an aggregate level, the supply chain responsiveness and the bottleneck capacity are determined based on the average lead time duration and the production capacity utilization of the different processes. In this paper, we integrate the impact of the production capacity on the lead time durations by the use of queuing networks. We add a supply chain responsiveness and production capacity performance measure on top of an inventory performance measure into the guaranteed service approach ([Graves and Willems, 2000]). Such an integrated approach allows us, for example, to measure the impact of a higher demand volume on the production capacity utilization, the lead times and the strategic stock.

The model described in this paper is part of a larger research program on end-to-end vaccine supply chain design. Therefore, in the intent to position this paper precisely, we refer to a generic approach which boils down to a five step framework ([Decouttere et al., 2015]), where the supply chain design problem is constructed through a scenario approach. A list of scenarios is constructed along the possible values of particular design and event parameters. In this way, at maximum, the total number of scenarios equals the full factorial design of these design and event parameters. This helps to explain why we do not include all features in the model described in this paper. It is to be considered as a building block. Figure 1 visualizes the approach. Each scenario, being one realization of each design parameter and each event parameter, can be visualized by the dotted rectangle. On the left we have the scenario description which constitutes out of the design parameter values, the event parameter values, other relevant scenario parameters (common to all designs and all events) and decision variables mapping the degree of freedom within the scenario itself which will be subject to optimization. In the middle of the figure it is shown that a scenario is built through three distinct models: a flow model which models the operational issues, a financial model which models the financial side of the scenario and finally a value model which models the value based issues. These are interconnected, dependent and lead together to the calculation of Key Performance Indicators (KPIs), which may be operational, financial or value based. This model based part of the scenario may be complemented by some non-model based scenario characteristics: KPIs which come straight from the scenario description. This paper describes a flow model that covers a part of the

Figure 1: Position of this paper



supply chain and turns a subset of the scenario description parameters into a couple of operational KPIs. We obtain the production capacity utilization, the responsiveness and the optimal total stock, which are the sum of the safety stocks and early arrival stocks. These concepts will be defined in more detail later in this paper.

The remainder of this paper is organized as follows. Section 2 introduces the guaranteed service approach, the stochastic service approach, the hybrid service approach and the related literature. Section 3 extends the guaranteed service approach methodology and the integration of production capacity into this approach. In Section 4 we show our results for an extended data set. The real-life applicability of our work to the rotavirus vaccine supply chain is shown in Section 5. In Section 6, we draw conclusions and identify implications for future research.

2 Literature Review

2.1 Guaranteed Service Approach

[Graves and Willems, 2003] distinguish the GSA and the Stochastic Service Approach as two main approaches in the literature on inventory models for multi-echelon supply chains. The GSA optimizes the strategic safety stock placement in supply chains while providing a high customer service level. This work is relevant for managers who face the pressure of reducing inventories in an existing supply chain as the production capacity and the lead times are fixed. A recent comprehensive survey of [Eruguz et al., 2016] classifies the existing Guaranteed Service Approach (GSA) literature along three dimensions: the considered assumptions, the developed solution methods and the industrial application. The GSA assumes that every supply chain stage quotes an outbound guaranteed service time: this is the time by which the stage under consideration satisfies the (internal or external) demand of the next stage. This

quoted guaranteed service time complies with 100% service to its customers. The GSA assumes that safety stock can be used to cope with normal demand variability and other additional countermeasures beyond safety stock must be used when the demand exceeds a normal variability level. In the GSA, such available countermeasures, for example accelerated production and overtime, refer to the operating flexibility assumption of the supply chain stages required to ensure a 100% delivery service. The assumption to be able to cope with normal demand variability is often referred to as *the assumption of bounded demand*. However, the external demand may not be bounded. This implies that one portion of the demand will be buffered with stock, another portion will be handled with operating flexibility measures and the remaining portion will be backlogged or lost. To perform such flexibility measures, excess production capacity may need to exist in the supply chain. Traditionally, the external demand is assumed to be stationary and is propagated to all the other stages. [Schoenmeyr and Graves, 2009] demonstrate the benefits of including the demand forecasting process for serial and assembly supply chains. We also refer the interested reader to [Graves and Willems, 2008] and [Neale and Willems, 2009] as they show how to propagate non-stationary demand into guaranteed service supply chains.

Another assumption is that each stage operates under a periodic review base stock policy with a common review period. The GSA has been extended toward different inventory policies: optimal batch ordering policies, optimal continuous review policies and more general (stage-dependent) review policies have been studied by [Li et al., 2013], [Chen and Li, 2015], [Bossert and Willems, 2007], respectively for different supply chain topologies. [Eruguz et al., 2014] optimize nested power-of-two reorder intervals and order-up-to levels in guaranteed service supply chains simultaneously. [Klosterhalfen et al., 2014] develop an exact approach to integrate static dual supply in the GSA by adopting an order-splitting policy among the suppliers.

[Graves and Willems, 2003, Graves and Willems, 2005] and [Moncayo-Martínez and Zhang, 2013] combine the GSA with supply chain configuration: some stages face the selection of an option of the functionality of the stage. These options may differ in its direct cost and lead time. [Moncayo-Martínez and Zhang, 2013] minimize the total supply chain cost and the supply chain's responsiveness of a supply chain configuration problem simultaneously using a bi-objective MAX-MIN ant system. [Funaki, 2012] and [You and Grossmann, 2010, You and Grossmann, 2011a] combine the GSA with facility location decisions: production or distribution facilities have to be located and the appropriate stock levels have to be set to optimize the total supply chain costs and/or supply chain's responsiveness. For these works the production capacity remains fixed.

[Lesnaia, 2004] shows that the optimization of strategic safety stock levels with the GSA is a NP-hard problem for general acyclic networks. Dynamic programming algorithms have been developed for different supply chain network topologies: for serial supply chains ([Chen and Li, 2015]), assembly type networks ([Funaki, 2012]), spanning trees ([Graves and Willems, 2000], [Graves and Willems, 2005]), networks with clusters of commonality ([Humair and Willems, 2006]) and general acyclic networks with a generalized cost function ([Humair and Willems, 2011]). [Magnanti et al., 2006] use successive piecewise linearization to approximate the nonlinear (concave) cost objective function and add redundant constraints to improve the computation time for the optimization of the strategic safety stock in general networks. Two simple heuristics are developed by [Shu and Karimi, 2009]. These heuristics use iterative linear approximations of the objective function to solve the strategic safety stock problem for general acyclic networks. [Li and Womer, 2008] and [Li and Jiang, 2012] formulate the earlier mentioned supply chain configuration problem and the strategic safety stock problem in general acyclic networks as equivalent project scheduling problems. These papers show how temporal constraints and resource constraints can be imposed by using the project scheduling representations. The authors propose constraint programming based solution approaches. [Grahl et al., 2014] relax the assumption of identical guaranteed service times toward all the successors of a particular stage. As the complexity of the problem increases by relaxing this assumption, the authors apply metaheuristics to solve it.

2.2 Stochastic Service Approach

[Graves and Willems, 2003] also explain the Stochastic Service Approach (SSA). Both GSA and SSA are used to solve the same multi-echelon inventory problem, but have different assumptions concerning the flexibility of the supply chain stages and the role of safety stock. The SSA relaxes the strict service guarantee (100% delivery service) and assumes that the service of each stage is variable: it allocates the stock to different stages by assuming that occasional upstream shortages may cause delivery delays. Furthermore, this approach assumes that safety stock is the only means to deal with variability. This implies that the supply chain is seen as inflexible and the production capacity is fixed. [Klosterhalfen and

Minner, 2010] propose a simulation study to compare the resulting costs of the GSA and the SSA for a system with one warehouse and multiple retailers. The authors assign additional costs for using operating flexibility in the GSA and conclude that the GSA performs better in case of moderate operating flexibility costs, long warehouse processing times and high retailer service levels. The computational attractiveness of the GSA over SSA is confirmed by this research. [Graves and Willems, 2003] also combine the SSA with the supply configuration problem mentioned earlier.

2.3 Hybrid Service Approach

[Klosterhalfen et al., 2013] integrate both the GSA and the SSA into supply chain multi-echelon inventory optimization to benefit from the advantages of both approaches. The authors propose a Hybrid Service Approach (HSA), instead of considering the GSA and SSA as mutually exclusive frameworks, which determines the best performing approach (GSA or SSA) for each supply chain stage by partitioning the serial supply chain into subnetworks. The numerical study of this work concludes that a pure GSA or SSA is dominated by the HSA in terms of total inventory costs, i.e. pipeline and on-hand stock. The authors mention two interesting future research areas: (1) the integration of capacity constraints and (2) the extension of the HSA to other network structures. As motivated in the next subsection, we elaborate on the integration of capacity constraints into the GSA.

2.4 Our approach

In this paper, we further elaborate on the GSA as we observed that this approach is tractable from both a modeling and a computational point of view. [Graves and Willems, 2003] describe that characterizing the replenishment lead time is extremely challenging for the SSA: for each stage, the replenishment lead time is variable and equals the sum of its deterministic processing time and the variable delay of its suppliers. For one stage, the number of combinations of predecessors that can be out of stock increases in a combinatorial way with the number of predecessors. Furthermore, a supplier's supplier might be out of stock in a multi-echelon inventory system. [Chen and Li, 2015] confirm that the GSA has the advantage that the inventory optimization problem can be formulated as a deterministic programming problem rather than a stochastic programming problem. As mentioned earlier, the GSA has been combined with supply chain configuration and supply chain design decisions. [Humair et al., 2013] confirm the savings for several real-world companies with inventory reductions by implementing the GSA. [Billington et al., 2004], [Farasyn et al., 2011], [Wieland et al., 2012] and [Moncayo-Martínez et al., 2014] elaborate on the GSA as a foundation of a multi-echelon inventory tool to manage inventories at Hewlett-Packard, Procter & Gamble, Intel and a company in the automotive industry respectively.

A large body of the publications in this literature review focuses on inventory optimization or minimizing total supply chain costs and assumes infinite production capacity. In this work the assumption of infinite production capacity will be relaxed and we consider the production capacity utilization and supply chain responsiveness as additional KPIs. [Sitompul et al., 2008] emphasize that locating strategic safety stock becomes a lot more complicated if production capacity constraints are taken into account. For serial supply chains, the authors derive the safety stock and excess production capacity based on the available production capacity and standard deviation of the demand during the net replenishment lead time. The authors use simulation to estimate a correction factor to obtain the same stock-out probability as in the uncapacitated case. [Graves and Schoenmeyr, 2016] published an extension of the GSA model ([Graves and Willems, 2000]) to include capacity constraints and analytically characterized the necessary base stock levels. The authors show how existing dynamic programming algorithms for the uncapacitated case can be adapted to the capacitated case. Production capacity constraints have also been integrated in guaranteed service supply chains by [Jung et al., 2008]. Their mathematical program models the production quantities and the excess production capacity as decision variables and minimizes inventory holding costs and backordering costs. Our approach focuses on the integration of production capacity into guaranteed service supply chains by the use of queuing networks. This approach allows us to calculate lead times which depend on the production capacity.

3 Methodology

3.1 Guaranteed Service Approach

As in [Humair et al., 2013], we further build on the distinction between the guaranteed service approach with deterministic lead times (GSA-DET) and the guaranteed service approach with variable lead times (GSA-VAR). The lead time of a stage represents the time from the availability of all the inputs of this stage until the output is ready to serve the (internal or external) demand of the next stage and may include material handling, machine processing, transportation time and waiting time, but also time to undergo regulatory, quality and release procedures. In GSA-DET, the value for these lead times is deterministic and two easy heuristics exist to choose this value in case of an empirical distribution of the lead times: (1) fixing the lead times to their mean value and (2) fixing the lead times to their maximum value. However, such a reasonable heuristic leads to inaccurate inventory levels and an inaccurate responsiveness performance measure. To the best of our knowledge, the work of [Humair et al., 2013] and [Neale and Willems, 2009] are the only manuscripts that allow lead time variability into the GSA. These authors demonstrate how inventory levels can be determined in a more accurate way instead of using these two heuristics.

The remainder of this section is organized as follows. First, we introduce the modeling framework of the GSA with deterministic lead times. We refer the interested reader to [Graves and Willems, 2000] for a more detailed description of GSA-DET. Next, we present how [Humair et al., 2013] allow for lead-time variability. Finally, we show our approach which integrates production capacity by the use of queuing networks and still allows for demand and lead time variability.

3.2 Guaranteed Service Approach with Deterministic Lead Times (GSA-DET)

According to [Graves and Willems, 2000], a supply chain stage represents a processing resource in the supply chain and might be the procurement of raw materials, production of components or subassemblies, production of assemblies and testing of the finished goods or the transportation from a distribution center to a regional warehouse. The supply chain network can also be represented by a graph where the nodes correspond to the supply chain stages and the arcs denote the precedence relationships between the nodes. A strategic safety stock point and its associated volume can be located at each stage.

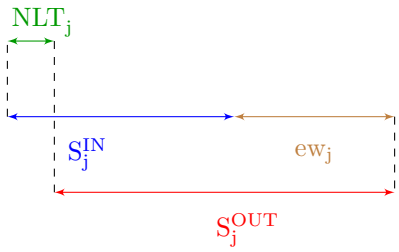
For GSA-DET, we assume that each stage j has a *deterministic production lead time*, $\mathrm{ew_j}$. This means, for example, if the inputs for stage j are available at time t, then the considered stage has completed its processing request at time $t + \mathrm{ew_j}$. [You and Grossmann, 2010, You and Grossmann, 2011a] use the term *order processing time* instead of production lead time and assume the order processing time to be independent of the order size. [Graves and Willems, 2000] mention the important assumption that the production lead time is not influenced by the capacity utilization. We will relax this assumption in our model.

The GSA assumes that each supply chain stage quotes an outbound guaranteed service time ($\mathrm{S_j^{OUT}}$). The $\mathrm{S_j^{OUT}}$ is the time by which the stage can guarantee a 100 % delivery service. The stage must hold sufficient inventory such that it is able to ensure the strict service guarantee. For each stage j, also an inbound guaranteed service time ($\mathrm{S_j^{IN}}$) is quoted. The $\mathrm{S_j^{IN}}$ is the time when all inputs of the stage are available and processing can start. Remark that the $\mathrm{S_j^{IN}}$ is equal to the maximum $\mathrm{S_j^{OUT}}$ of the independently preceding stages in case of a non-serial supply chain (e.g. assembly type supply chain topology). According to [Graves and Willems, 2000], this can be expressed as

$$\mathrm{S_j^{IN}} = \max_{i:(i,j)\in A} \{\mathrm{S_i^{OUT}}\} \tag{1}$$

where A denotes the set of arcs in the supply chain network. For every arc (i,j) $\in$ A, stage i replenishes its downstream stage j. For a stage j, GSA-DET assumes a deterministic replenishment lead time as both its guaranteed service times and the production lead time are deterministic. [You and Grossmann, 2010, You and Grossmann, 2011a] emphasize the replenishment lead time as an important difference between single-echelon and multi-echelon inventory systems. In a single stage inventory system, replenishment lead time is often exogenous and treated as parameter whilst in a multi-echelon inventory system it depends on the inventory level of the predecessor(s). We denote the replenishment time as the *Net (replenishment) Lead Time* (NLT). The GSA computes the NLT as the sum of the stage's inbound

Figure 2: Calculation of the net lead time (based on [You and Grossmann, 2010])



guaranteed service time and the production lead time minus the outbound guaranteed service time, or formally:

$$NLT_j = S_j^{IN} + ew_j - S_j^{OUT} \tag{2}$$

Figure 2 visualizes this equation. As explained in detail in [You and Grossmann, 2010], we highlight two extreme cases regarding equation (2):

- $S_j^{OUT} = 0$, i.e. $NLT_j = S_j^{IN} + ew_j$. An outbound guaranteed service time equal to zero means that the stage can fulfill the order of the downstream stage immediately. This requires that the stage needs to hold the most strategic safety stock. In such case, a stage is denoted as operating in *push* mode.
- $S_j^{OUT} = S_j^{IN} + ew_j$, i.e. $NLT_j = 0$. The reception of the necessary inputs of the predecessors and the production lead time is within the quoted outbound guaranteed service time. This implies that the stage does not need to hold any strategic safety stock. In such case, a stage is denoted as operating in *pull* mode.

[You and Grossmann, 2011a] show that the supply chain's responsiveness can be measured as the S_j^{OUT} of the last stage of the supply chain (e.g. customer markets). This responsiveness measure quantifies the maximum time within which the (external) demand is satisfied. In case of multiple customer markets, the supply chain's responsiveness may be computed as (1) the maximum S_j^{OUT} of the markets or (2) a weighted average value of each customer market's S_j^{OUT}. In such a case we will opt for the first method.

[Graves and Willems, 2003] derive the strategic Safety Stock (SS_j) and Base Stock (BS_j) level held at the stage under consideration as a function of its net replenishment lead time:

$$SS_j = k_j \sigma_j^D \sqrt{NLT_j} \tag{3}$$

$$BS_j = \mu_j^D NLT_j + k_j \sigma_j^D \sqrt{NLT_j} \tag{4}$$

These results rely on a common way to set a node's demand bound, namely $\mu_j^D NLT_j + k_j \sigma_j^D \sqrt{NLT_j}$. In these equations, k_j represents the *safety stock factor* at stage j. This parameter determines the stage's demand that can be covered with the stage's safety stock. A common assumption is that the demand of stage j is normally distributed with mean μ_j^D and standard deviation σ_j^D. The expected *Work-In-Process* inventories (WIP) or *pipeline stock* at stage j can be determined using Little's law ([Little, 1961]):

$$WIP_j = ew_j \mu_j^D \tag{5}$$

We formulate the complete deterministic mathematical program (denoted as problem P1) of the guaranteed service framework:

$$\min \quad \sum_{j \in \mathrm{J}} \mathrm{k_j} \sigma_{\mathrm{j}}^{\mathrm{D}} \sqrt{\mathrm{S_j^{IN}} + \mathrm{ew_j} - \mathrm{S_j^{OUT}}} \tag{6}$$

$$\text{subject to} \quad \mathrm{S_j^{OUT}} - \mathrm{S_j^{IN}} \leq \mathrm{ew_j} \qquad \forall j \in \mathrm{J} \tag{7}$$

$$\mathrm{S_j^{IN}} - \mathrm{S_i^{OUT}} \geq 0 \qquad \forall (i,j) \in \mathrm{A} \tag{8}$$

$$\mathrm{S_j^{OUT}} \leq \mathrm{s_j} \qquad \forall j \in \mathrm{J_E} \tag{9}$$

$$\mathrm{S_j^{IN}}, \mathrm{S_j^{OUT}} \geq 0 \qquad \forall j \in \mathrm{J} \tag{10}$$

The objective function (6) minimizes the strategic SS across the supply chain. Remark that the strategic SS levels are nonlinearly related to the NLTs. Constraint set (7) implies that the NLTs are positive. Constraints (8) are a linearization of (1): the stage's $\mathrm{S_j^{IN}}$ is the time when the outputs of all the preceding stages are available. The maximal quotable $\mathrm{S_j^{OUT}}$ to an end node ($j \in \mathrm{J_E}$) in the supply chain is denoted by $\mathrm{s_j}$. Constraints (9) enforce that the outbound guaranteed service times to the end nodes satisfy the exogenously quoted guaranteed service times. In case of multiple end nodes, our aggregate supply chain responsiveness measure (R) is defined as the maximum of the outbound guaranteed service times to the end nodes. The final set of constraints (10) assures the nonnegativity of the guaranteed service times.

3.3 Guaranteed Service Approach with Variable Lead Times (GSA-VAR)

[Humair et al., 2013] emphasize the importance of integrating variable lead times into the GSA as every supply chain is confronted with variable lead times. The authors develop a closed-form expression to compute the SS for the GSA with variable lead times such that the total inventory cost can be minimized. They provide a numerical comparison of SS and WIP levels of GSA-VAR and two heuristics for fixing the lead times of an empirical distribution for GSA-DET: (1) fixing the lead times to their mean value and (2) fixing the lead times to their maximum value. Compared to GSA-VAR, the SS levels may be significantly underestimated for both heuristics. The WIP levels are equal for both GSA-VAR and GSA-DET in case the lead times are fixed to their mean value, but they are overestimated in case the lead times are fixed to their maximum value.

We note that in case of variable lead times, processing at a stage might be finished early because of shorter lead time realizations. In this case, it is impossible to pass stock to the next stage as the GSA assumes that this downstream stage can only start processing at its $\mathrm{S_j^{IN}}$. This leads to *Early-Arrival Stock* (EAS) in the supply chain. Such a phenomenon only occurs if the lead time realization is smaller than the difference between the outbound and inbound guaranteed service time of the stage. In this case, the NLT is negative and this leads to a *negative shortfall*. That is, the shortfall is negative when a stage has replenished all the demand that it has filled and also some demand it has yet to fill. For completeness, we note that there is a *positive shortfall* when the NLT is positive and no shortfall when the NLT equals zero. The shortfall is positive when the demand associated with the replenishment order has not yet been filled.

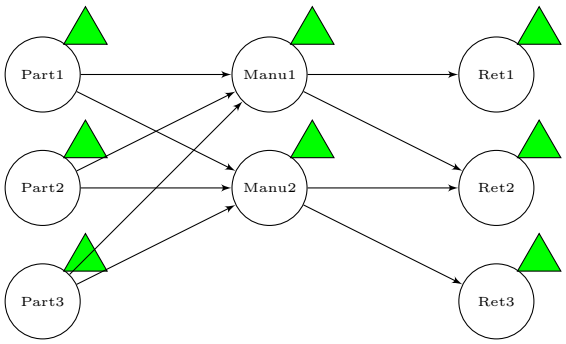
[Humair et al., 2013] determine a closed-form expression for the computation of the SS (for GSA-DET, see (3)) and the average EAS in case of variable lead times. In this work, we show the resulting closed-form expressions. We refer the interested reader to the appendix of [Humair et al., 2013] for the full derivation of these equations:

$$\mathrm{SS_j} = \mathrm{k_j} \sqrt{\mathrm{Y(T)}(\sigma_{\mathrm{j}}^{\mathrm{D}})^2 + (\mu_{\mathrm{j}}^{\mathrm{D}})^2 \mathrm{Z(T)}} \tag{11}$$

$$\mathrm{EAS_j} = \mu_{\mathrm{j}}^{\mathrm{D}} (\mathrm{Y(T)} - \mathrm{ew_j} + \mathrm{T}) \tag{12}$$

where $\mathrm{T} = \mathrm{S_j^{OUT}} - \mathrm{S_j^{IN}}$ and Y() and Z() are functions that depend on the probability distribution of the lead time and are specified in Appendix A (based on [Humair et al., 2013]). The expected value and the variance of the NLT conditional on being positive are computed by Y(T) and Z(T) respectively. In equation (11), the square root term denotes the standard deviation of the positive shortfall. In equation (12), $\mathrm{ew_j}$ represents the mean value of the general lead time distribution of stage j. Remark that the closed-form expressions hold for both discrete and continuous distributions of the lead times. The total stock for GSA-VAR can be determined by replacing the objective function (6) of problem P1 by the sum of (11) and (12) over all nodes.

Figure 3: Structure of supply chain Chain 01 of [Willems, 2008]



3.4 Incorporation of production capacity into the Guaranteed Service Approach with Variable Lead Times

Figure 3 shows the supply chain structure of Chain 01 of [Willems, 2008] consisting of three echelons and eight stages: three part suppliers (Part1, Part2 and Part3), two manufacturers (Manu1 and Manu2) and three retailers (Ret1, Ret2 and Ret3). The triangles represent candidate strategic SS locations. For the ease of explanation, we refer to the stages as nodes, denote the set of nodes as J and partition the set of nodes according to the three echelons: J_P, J_M and J_R where J_P is the set of part supplier nodes, J_M is the set of manufacturer nodes and J_R is the set of retailer nodes. At each node, an activity with a certain duration has to be performed. The arcs represent the precedence relationships between the nodes and we assume an assembly type supply chain: the activities at Manu1, Manu2 and Ret2 can only start when all the predecessors are ready. Remark that such an activity may include machine processing time, but also waiting, material handling, transportation time or time to undergo regulatory and quality procedures. We refer to the node's total (average) activity duration as the node's average lead time.

Furthermore we assume that the average lead time and its variability are available for the part supplier and the retailer echelon. The manufacturing echelon consists of high-speed processing nodes. Remark that such a node may represent a complex production system and could be decomposed into a subnetwork of nodes. For each manufacturing node we model multiple, identical production lines and assume that their unit processing time and variability are known and the lead time and its variability are to be determined. One of our contributions is to model these lead times as a function of the installed production capacity for guaranteed service supply chains. In case of large volumes, inducing a high capacity utilization, the waiting time for such a processing activity may increase drastically. Such a long waiting time has a negative impact on the corresponding node's outbound service time and subsequently damages the supply chain's responsiveness if no additional production capacity and/or inventory buffers can be inserted.

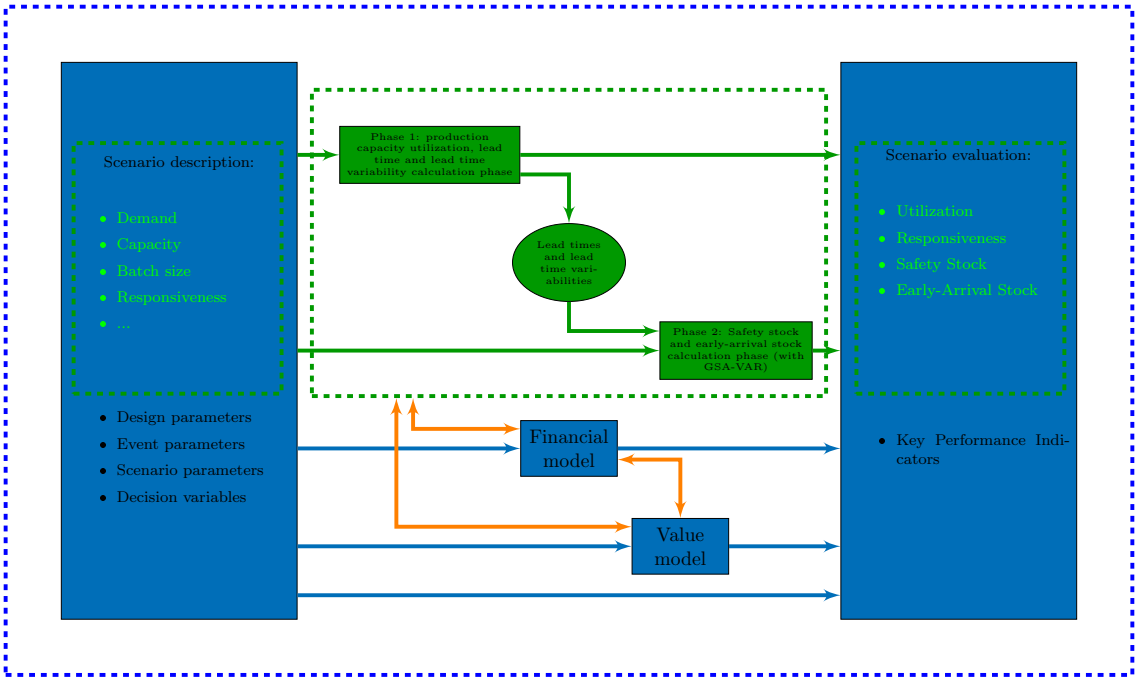
We show our approach in Figure 4 which further elaborates on the flow model in Figure 1. A subset of the parameters and decision variables of the scenario description is needed for the flow model (shown in green). The flow model consists of two phases which are visualized by green rectangles. The first phase decomposes the manufacturing nodes and calculates the production capacity utilization, lead time and lead time variability of the manufacturing nodes by the use of queuing networks. The production capacity utilization is one of the KPIs of our flow model and is shown on the right of this figure. The second phase uses the releveant information of the scenario description and the calculated lead times and lead time variabilities in the previous phase to calculate the supply chain's safety stock and early-arrival stock. The supply chain responsiveness is a KPI which comes straight from the scenario description.

3.4.1 Phase 1: Lead time and lead time variability calculation phase

The general outline of this phase is as follows:

1. We propagate the external demand and its variability (in units) to all the internal nodes in the SCN

Figure 4: Two phases for production capacity integration into GSA-VAR



2. For a manufacturing node, we set the arrival stream to the production line equal to the node's demand

3. We propagate the square coefficents of variation of the arrival processes downstream

4. At the manufacturing echelon we assume serial process batching and compute the manufacturing nodes' utilization, lead time and lead time variability

We assume that the average demand per time unit and its standard deviation, denoted as $\mu_{\mathrm{j}}^{\mathrm{D}}$ and $\sigma_{\mathrm{j}}^{\mathrm{D}}$ respectively, is available for $j \in \mathrm{J_R}$ and follows a stationary process. For an assembly type supply chain, the internal or dependent demand of the manufacturer echelon $\mathrm{J_M}$ and the part supplier echelon $\mathrm{J_P}$ can be calculated sequentially by:

$$\mu_{\mathrm{j}}^{\mathrm{D}} = \sum_{i:(j,i)\in A} \theta_{\mathrm{j,i}} \mu_{\mathrm{i}}^{\mathrm{D}} \tag{13}$$

where $\theta_{j,i}$ represents the number of items of upstream node j required for downstream node i according to the bill of material.

The total volume that needs to be available to satisfy the internal or external demand of the corresponding node j is denoted by λ_j and equals the node's demand ($\mu_{\mathrm{j}}^{\mathrm{D}}$) when there are no disposals caused by technical issues (e.g. yield, scrap or write-offs).

The standard deviation of the demand for the manufacturing nodes $j \in \mathrm{J_M}$ can be computed by:

$$\sigma_{\mathrm{j}}^{\mathrm{D}} = \sqrt{\sum_{\mathrm{i:(j,i)\in A}} \sum_{\mathrm{i':(j,i')\in A}} \theta_{\mathrm{j,i}} \theta_{\mathrm{j,i'}} \sigma_{\mathrm{i}}^{\mathrm{D}} \sigma_{\mathrm{i'}}^{\mathrm{D}} \phi_{i,i'}} \tag{14}$$

where i and i' each represent a retailer demand node and $\phi_{\mathrm{i,i'}}$ denotes the correlation between the retailer demand streams ($\phi_{\mathrm{i,i}} = 1$ by definition). Remark that the demand at internal stages may be correlated even when end-item demand streams are independent ([Humair and Willems, 2011]: online supplement). For the supply chain network displayed in Figure 3 the reader can check that the demand streams seen by the part supplier nodes are correlated. We use the following expression to calculate the standard deviation of the part nodes' demand ($\sigma_{\mathrm{p}}^{\mathrm{D}}$ with $p \in \mathrm{J_P}$) in terms of the standard deviation for the retailer demand nodes:

$$\sigma_{\mathrm{p}}^{\mathrm{D}} = \sqrt{\sum_{\mathrm{i}:(\mathrm{j},\mathrm{i})\in \mathrm{A}} \sum_{\mathrm{i}':(\mathrm{j},\mathrm{i}')\in \mathrm{A}} \widetilde{\theta}_{\mathrm{p,i}} \widetilde{\theta}_{\mathrm{p,i}'} \sigma_{\mathrm{i}}^{\mathrm{D}} \sigma_{\mathrm{i}'}^{\mathrm{D}} \phi_{i,i'}} \tag{15}$$

whereby $\widetilde{\theta}_{\mathrm{p,i}} = \sum_{\mathrm{j}:(\mathrm{p,j})\in \mathrm{A}} \theta_{\mathrm{p,j}} \theta_{\mathrm{j,i}}$ and $\widetilde{\theta}_{\mathrm{p,i}'} = \sum_{\mathrm{j}:(\mathrm{p,j})\in \mathrm{A}} \theta_{\mathrm{p,j}} \theta_{\mathrm{j,i}'}$. We refer the interested reader to the work of [You and Grossmann, 2011b] who use variance-to-mean ratios to propagate the demand variability. They show how risk pooling and the bullwhip effect can be modeled by using these ratios. In our work, we propagate supply variability by using Squared Coefficients of Variation (SCVs). We denote the SCVs of the unit arrival processes and unit service processes as $(\mathrm{C}_{\mathrm{j}}^{\mathrm{A}})^2$ and $(\mathrm{C}_{\mathrm{j}}^{\mathrm{S}})^2$ respectively and assume that the SCVs of the unit arrival processes are available for nodes $j \in \mathrm{J_P}$. For manufacturing nodes $j \in \mathrm{J_M}$, the SCVs of the unit service processes are also given. For the parts echelon, we assume that the SCVs of the departing processes, $(\mathrm{C}_{\mathrm{j}}^{\mathrm{D}})^2$, are equal to $(\mathrm{C}_{\mathrm{j}}^{\mathrm{A}})^2$ as this echelon does not contain machine processing operations.

For the manufacturing echelon, we calculate each node's lead time and corresponding variability. According to (11) and (12), this information influences the necessary SS levels and emerging EAS. We compute the expected lead time and its variability by taking production capacity constraints into account using a G/G/m queuing model. We allow the lead time and its variability to be dependent of the batch size and assume general interarrival and processing time distributions. We assume serial process batching: the units in the process batch are processed one by one but setup time is needed between two batches. Furthermore we allow for multiple identical production lines, denoted by m. For nodes $j \in \mathrm{J_M}$, we compute the expected lead time, $\mathrm{ew_j}$, based on [Lambrecht and Vandaele, 1996]:

$$\mathrm{ew_j} = \mathrm{ew_j^C} + \mathrm{ew_j^Q} + \mathrm{ew_j^P} \tag{16}$$

where $\mathrm{ew_j^C}$ is the batch collection time, $\mathrm{ew_j^Q}$ is the batch waiting time and $\mathrm{ew_j^P}$ is the batch processing time. We omit the batch collection time as it is less relevant for our industrial application. For $\mathrm{ew_j^P}$, we assume that the unit processing times are independent and identically distributed and a setup time τ_{j} which is independent of the processing times. The expected batch processing time can be computed similarly to [Lambrecht and Vandaele, 1996]:

$$\mathrm{ew_j^P} = \tau_{\mathrm{j}} + \mathrm{Q_j t_j} \tag{17}$$

where $\mathrm{Q_j}$ and $\mathrm{t_j}$ represent the batch size and the unit processing time respectively. The expected waiting time is derived by the use of the bulk arrival-bulk service approach. Then the following batch arrival process quantities can be characterized ([Lambrecht and Vandaele, 1996]):

$$\lambda_{\mathrm{j}}^{\mathrm{BA}} = \frac{\lambda_{\mathrm{j}}}{\mathrm{Q_j}} \tag{18}$$

$$(\mathrm{C}_{\mathrm{j}}^{\mathrm{BA}})^2 = \frac{(\mathrm{C}_{\mathrm{j}}^{\mathrm{A}})^2}{\mathrm{Q_j}} \tag{19}$$

where the superscript BA refers to the batch arrival process, $\lambda_{\mathrm{j}}^{\mathrm{BA}}$ and λ_j refer to the arrival rate of the batch process and the individual units respectively and $(\mathrm{C}_{\mathrm{j}}^{\mathrm{BA}})^2$ is the SCV of the batch arrival process. The batch service processes, denoted by the superscript BS, are described as follows:

$$(\sigma_{\mathrm{j}}^{\mathrm{BS}})^2 = (\sigma_{\mathrm{j}}^{\tau})^2 + \mathrm{Q_j}(\sigma_{\mathrm{j}}^{\mathrm{t}})^2 \tag{20}$$

$$(\mathrm{C}_{\mathrm{j}}^{\mathrm{BS}})^2 = \frac{(\sigma_{\mathrm{j}}^{\mathrm{BS}})^2}{(\mathrm{ew_j^P})^2} \tag{21}$$

where $(\sigma_{\mathrm{j}}^{\mathrm{BS}})^2$ and $(\mathrm{C}_{\mathrm{j}}^{\mathrm{BS}})^2$ is the variability and SCV of the batch service time for all j in $\mathrm{J_M}$ respectively. The standard deviation of the setup time and the unit processing times are denoted by $\sigma_{\mathrm{j}}^{\tau}$ and $\sigma_{\mathrm{j}}^{\mathrm{t}}$. For nodes $j \in \mathrm{J_M}$, the utilization rate ρ_j is equal to:

$$\rho_{\mathrm{j}} = \frac{\lambda_{\mathrm{j}}^{\mathrm{BA}} \mathrm{ew}_{\mathrm{j}}^{\mathrm{P}}}{\mathrm{m}} < 1 \tag{22}$$

The utilization rate must be strictly smaller than 100% to avoid the node's explosion. Indeed, the suggested approximation of [Whitt, 1983] shows that the expected waiting time increases in a nonlinear way when the utilization rate increases for a batch arrival - batch service G/G/m queue:

$$\mathrm{ew}_{\mathrm{j}}^{\mathrm{Q}} = \omega\left(\rho_{\mathrm{j}}, (\mathrm{C}_{\mathrm{j}}^{\mathrm{BA}})^2, (\mathrm{C}_{\mathrm{j}}^{\mathrm{BS}}), \mathrm{m}\right) \left(\frac{(\mathrm{C}_{\mathrm{j}}^{\mathrm{BA}})^2 + (\mathrm{C}_{\mathrm{j}}^{\mathrm{BS}})^2}{2}\right) \left(\frac{\rho_{\mathrm{j}}^{\sqrt{2(\mathrm{m}+1)}-1}}{\mathrm{m}(1-\rho_{\mathrm{j}})}\right) \mathrm{ew}_{\mathrm{j}}^{\mathrm{P}} \tag{23}$$

where the correction factor ω applies to a node under heavy traffic conditions with multiple parallel production lines ([Whitt, 1993]). Note that we use the Kraemer and Langenbach-Belz formula ([Kraemer and Langenbach-Belz, 1976, Lambrecht et al., 1998]) in case $m = 1$. An expression for $(\mathrm{C}_{\mathrm{j}}^{\mathrm{BA}})^2$ with $j \in \mathrm{J_M}$ is less trivial because the arrival stream at manufacturing node j depends on the departing streams from its predecessors. To obtain this aggregate SCV we use the approximation as in [Whitt, 1983]. Then, the batch departure SCV of each manufacturing node can be determined by the following approximation ([Whitt, 1983]):

$$(\mathrm{C}_{\mathrm{j}}^{\mathrm{BD}})^2 = 1 + (1-\rho_{\mathrm{j}}^2)((\mathrm{C}_{\mathrm{j}}^{\mathrm{BA}})^2 - 1) + \frac{\rho_{\mathrm{j}}^2}{\sqrt{\mathrm{m}}}((\mathrm{C}_{\mathrm{j}}^{\mathrm{BS}})^2 - 1) \tag{24}$$

Finally, for nodes $j \in \mathrm{J_M}$, the expected lead time, ew_{j}, is the sum of the expected batch waiting time and the batch processing time. The lead time variability can be obtained by:

$$(\sigma_{\mathrm{j}}^{\mathrm{ew}})^2 = (\sigma_{\mathrm{j}}^{\mathrm{ew}^{\mathrm{Q}}})^2 + (\sigma_{\mathrm{j}}^{\mathrm{ew}^{\mathrm{P}}})^2 + 2\mathrm{cov}(\mathrm{ew}_{\mathrm{j}}^{\mathrm{Q}}, \mathrm{ew}_{\mathrm{j}}^{\mathrm{P}}) \tag{25}$$

which includes the variance of the expected waiting time, the variance of the batch processing time and their covariance as they are not independent. Based on simulation results, [Lambrecht and Vandaele, 1996] postulate that such covariance is hard to obtain in case of a lot sizing model and drop this computation as they expect it to have a low contribution to $(\sigma_{\mathrm{j}}^{\mathrm{ew}})^2$. In this work, we also consider the computation of this covariance term as out of scope. We opt for the approximation of [Whitt, 1993] to derive the variance of the batch waiting time and refer to Appendix B for more detail.

All the previous information allows us to compute the first and second order moment of the lead time distribution. However, for the intermediary computations for (11) and (12) the lead time probability distribution should be known. Even when the probability distribution of the waiting time and the machine processing time are known, the probability distribution of the lead time might be hard to obtain. Experiments of [Lambrecht et al., 1998], [Lambrecht and Vandaele, 1996] and [Vandaele, 1996] indicate that the (right-) skewed lognormal distribution provides a good fit for the lead time distribution function. For that reason we will also use a lognormal distribution and determine its scale parameter $\delta_{\mathrm{j}} \in [-\infty, \infty]$ and shape parameter $\gamma_{\mathrm{j}} > 0$. Furthermore, this distribution does not require an additional parameter to shift the domain of the distribution. The authors above obtain δ_{j} and the γ_{j} of the lead time distribution by:

$$\delta_{\mathrm{j}} = \ln\left(\frac{\mathrm{ew}_{\mathrm{j}}}{\sqrt{(\sigma_{\mathrm{j}}^{\mathrm{ew}})^2/(\mathrm{ew}_{\mathrm{j}})^2 + 1}}\right) \tag{26}$$

$$\gamma_{\mathrm{j}}^2 = \ln\left(\frac{(\sigma_{\mathrm{j}}^{\mathrm{ew}})^2}{(\mathrm{ew}_{\mathrm{j}})^2} + 1\right) \tag{27}$$

The supply chain SS and EAS can now be computed in the second phase. Note that both SS and EAS, for nodes $j \in \mathrm{J_M}$, now depend on the production capacity and its utilization.

3.4.2 Phase 2: Safety stock and early-arrival stock calculation phase with GSA-VAR

Phase 2 optimizes the supply chain's total stock under production capacity constraints. However, the supply chain's total inventories is a nonlinear function of the decision variables. Therefore we integrate a piecewise linear approximation formulation into GSA-VAR. This formulation is also referred to as the Multiple Choice Model and is based on the work of [Magnanti et al., 2006] and [Croxton et al., 2003]. We prefer the use of such a formulation as it does not require a specific supply chain topology, but the supply chain network cannot contain a directed cycle. The complete model formulation is now as follows:

$$\min \quad \sum_{j\in \mathrm{J}} \sum_{r\in \mathrm{R}} \beta_{\mathrm{j}}^{\mathrm{r}} \mathrm{y}_{\mathrm{j}}^{\mathrm{r}} + \alpha_{\mathrm{j}}^{\mathrm{r}} \mathrm{z}_{\mathrm{j}}^{\mathrm{r}} \tag{28}$$

$$\text{subject to} \quad \sum_{r\in \mathrm{R}} \mathrm{z}_{\mathrm{j}}^{\mathrm{r}} = \mathrm{S}_{\mathrm{j}}^{\mathrm{OUT}} - \mathrm{S}_{\mathrm{j}}^{\mathrm{IN}} \qquad \forall j \in \mathrm{J} \tag{29}$$

$$\mathrm{S}_{\mathrm{j}}^{\mathrm{OUT}} \leq \mathrm{s}_{\mathrm{j}} \qquad \forall j = \mathrm{J}_{\mathrm{E}} \tag{30}$$

$$\mathrm{M}_{\mathrm{j}}^{\mathrm{r}-1} \mathrm{y}_{\mathrm{j}}^{\mathrm{r}} \leq \mathrm{z}_{\mathrm{j}}^{\mathrm{r}} \leq \mathrm{M}_{\mathrm{j}}^{\mathrm{r}} \mathrm{y}_{\mathrm{j}}^{\mathrm{r}} \qquad \forall j \in \mathrm{J}, \forall r \in \mathrm{R} \tag{31}$$

$$\sum_{r\in \mathrm{R}} \mathrm{y}_{\mathrm{j}}^{\mathrm{r}} \leq 1 \qquad \forall j \in \mathrm{J} \tag{32}$$

$$\mathrm{S}_{\mathrm{j}}^{\mathrm{IN}} - \mathrm{S}_{\mathrm{i}}^{\mathrm{OUT}} \geq 0 \qquad \forall (i,j) \in \mathrm{A} \tag{33}$$

$$\mathrm{S}_{\mathrm{j}}^{\mathrm{IN}}, \mathrm{S}_{\mathrm{j}}^{\mathrm{OUT}} \geq 0 \qquad \forall j \in \mathrm{J} \tag{34}$$

$$\mathrm{z}_{\mathrm{j}}^{\mathrm{r}} \geq 0 \qquad \forall j \in \mathrm{J}, \forall r \in \mathrm{R} \tag{35}$$

$$\mathrm{y}_{\mathrm{j}}^{\mathrm{r}} \in \{0,1\} \qquad \forall j \in \mathrm{J}, \forall r \in \mathrm{R} \tag{36}$$

In this formulation, index $r \in \mathrm{R}$ denotes a finite number of intervals for each node's inventory objective function. Each interval is bounded by a lower and upper bound $\mathrm{M}_{\mathrm{j}}^{\mathrm{r}-1}$ and $\mathrm{M}_{\mathrm{j}}^{\mathrm{r}}$. For each interval, the node's objective function has a slope $\alpha_{\mathrm{j}}^{\mathrm{r}}$ and an intercept $\beta_{\mathrm{j}}^{\mathrm{r}}$. We denote $\mathrm{S}_{\mathrm{j}}^{\mathrm{OUT}} - \mathrm{S}_{\mathrm{j}}^{\mathrm{IN}}$ as the node's *value*. Remark that the node's objective function value contains the SS and EAS (equations (11) and (12)) and is a function of $\mathrm{S}_{\mathrm{j}}^{\mathrm{OUT}} - \mathrm{S}_{\mathrm{j}}^{\mathrm{IN}}$. We introduce the variable $\mathrm{z}_{\mathrm{j}}^{\mathrm{r}}$ which equals the node's value in interval r. Constraint (29) ensures that the node's value is equal to $\mathrm{S}_{\mathrm{j}}^{\mathrm{OUT}} - \mathrm{S}_{\mathrm{j}}^{\mathrm{IN}}$. For a $\mathrm{z}_{\mathrm{j}}^{\mathrm{r}} > 0$, this requires $\mathrm{y}_{\mathrm{j}}^{\mathrm{r}} = 1$ and $\mathrm{y}_{\mathrm{j}}^{\mathrm{r}} = 0$ otherwise. For each node, constraints (32) and (36) ensure that at most one $\mathrm{y}_{\mathrm{j}}^{\mathrm{r}}$ equals one. Finally, constraint set (31) enforces that each node's value is between their lower and upper interval bounds.

4 Experimental results with data from literature

The data set of [Willems, 2008] documents 38 multi-echelon supply chains that have been implemented in practice and has been used as a test bed by several manuscripts which propose a variant or a solution algorithm for GSA-DET. This data set classifies each stage according to one of these five classification labels: distribution, part supply, manufacturing, retail and transportation. [Humair et al., 2013] solve 26 of the 38 supply chains of this publicly available data set. These 26 supply chains employ variable lead times and [Humair et al., 2013] summarize the results for 12 supply chains. In this section we show our results for these same 12 supply chains. However, the data regarding the production capacity are unavailable for these supply chains. As these data are key for our contribution, we show how we use our own extension to the data set of [Willems, 2008].

4.1 Instance generation

For the instance generation, we use the supply chain topologies and following data from [Willems, 2008]:

- The average lead time and lead time standard deviation of the non-manufacturing stages
- The demand and demand standard deviation of the demand stages
- The service level of the demand stages
- The maximum service time which is the maximum time that the customer is willing to wait

For the manufacturing stages, $j \in J_M$, we use a random number generator to construct our own extended numerical data set:

- The SCVs of the unit service processes of the manufacturing echelon, $(C_j^S)^2$, are drawn from a continuous uniform distribution with range [0;1.5]
- The batch size Q_j is generated as a percentage of the corresponding node's demand. This percentage is drawn from a continuous uniform distribution with range [10%;50%].
- The setup time τ_j and unit processing time t_j are generated such that ρ_j <1. A percentage of the utilization rate dedicated to setup time and to actual processing time are drawn from a continuous uniform distribution with range [10%;20%] and [40%;80%] respectively.
- The setup time variance $(\sigma_j^\tau)^2$ is generated such that the corresponding SCV, $(C_j^\tau)^2$, follows a continuous uniform distribution with range [0;1.5]

Furthermore, without loss of generality, we assume that the number of parallel production lines, m, is equal to 1. As motivated earlier, we postulate a lognormal distribution for the lead time distribution function for the manufacturing stages. The determination of a lower bound on the number of intervals ($r \in R$) to guarantee a sufficiently precise approximation of a node's inventory performance measure is a hard problem. In combination with the generation of flow cover cuts, [Magnanti et al., 2006] show good results for general acyclic networks up to 100 stages while never requiring more than 12 intervals. As our largest supply chain contains 133 nodes (Chain 15), we set the number of intervals of one node to 20.

4.2 Results for 12 of the [Willems, 2008] supply chains

Both the mathematical programming model and the queuing networks of Section 3 are programmed in Microsoft Visual Studio 2010 and we use the callable library of IBM ILOG CPLEX 12.6 to solve the mathematical program. The code is executed on a Windows Server 2012 R2 with two Intel Xeon E5-2698 processors of 2.30 GHz and a RAM of 256 GB.

Table 1 shows the results of our approach (denoted as GSA-CAP) and those shown in [Humair et al., 2013]. For GSA-CAP, the number of stages with variable lead times will always include the number of manufacturing stages as our approach computes the lead times and their variability for these stages. The average utilization column represents the average of the production utilization rates of the manufacturing stages. Note that it is hard to compare the objective values of GSA-CAP and GSA-VAR as the computed lead time and lead time variability of the manufacturing nodes of GSA-CAP can be very different from the assumed lead time and lead time variability of GSA-VAR due to the random number realization to construct our own extended numerical data set. Our general findings are in line with those of [Humair et al., 2013]: the number of inventory locations increases compared to GSA-DET with a heuristic solution for obtaining the lead time. As in the case of Chain 03, the number of stages holding SS increases for GSA-CAP as the number of stages with variable lead times increases. The additional stages with variable lead times are the manufacturing stages which are now modeled with queuing networks to calculate their lead time and lead time variability. The increase in the number of SS locations is also in line with the intuition explained in [Humair et al., 2013]: supply chains that model variable lead times have more stages to hold SS to avoid EAS ramifactions as GSA-DET does not consider the latter type of stock.

4.3 Results for the impact of the production capacity integration into GSA-VAR

We further elaborate on Chain 01 of [Willems, 2008] and summarize the generated supply chain input data in Table 2. The table entries filled with a cross indicate data which are computed by our approach whilst the "NA"-entries show that the production capacity information is Not Applicable (NA) for non-manufacturing nodes.

Table 1: Results for 12 of the [Willems, 2008] supply chains

Chain name	Solution approach	Safety stock	Number of stages holding safety stock	Number of stages with variable lead times	Number of manufacturing stages	Average utilization	Total number of stages
01	GSA-VAR	8351	5	1	2	NA	8
01	GSA-CAP	8354	5	3	2	0.7327	8
03	GSA-VAR	15744	9	8	4	NA	17
03	GSA-CAP	13599	17	12	4	0.7996	17
05	GSA-VAR	1043257	25	16	5	NA	27
05	GSA-CAP	655769	26	17	5	0.7377	27
06	GSA-VAR	50062	16	16	10	NA	28
06	GSA-CAP	112399	26	22	10	0.7941	28
07	GSA-VAR	363	34	38	6	NA	38
07	GSA-CAP	273	38	38	6	0.7970	38
08	GSA-VAR	46030	5	23	4	NA	40
08	GSA-CAP	74118	38	25	4	0.8294	40
09	GSA-VAR	385497	35	11	4	NA	49
09	GSA-CAP	161922	44	15	4	0.7735	49
10	GSA-VAR	4890	55	21	13	NA	58
10	GSA-CAP	6060	58	22	13	0.7658	58
11	GSA-VAR	490	48	45	6	NA	68
11	GSA-CAP	408	63	57	6	0.7970	68
12	GSA-VAR	1511516	80	28	9	NA	88
12	GSA-CAP	732865	88	31	9	0.7354	88
14	GSA-VAR	6002	84	36	9	NA	116
14	GSA-CAP	15739	66	45	9	0.7671	116
15	GSA-VAR	775010	77	77	28	NA	133
15	GSA-CAP	455327	77	77	28	0.7674	133

Table 2: Summary of the input data for Chain 01

Node name	j	$(C_j^A)^2$	$(C_j^E)^2$	Q_j (units)	t_j (minutes)	τ_j (minutes)	ew_j (days)	σ_j^{ew} (daily)	μ_j^D (daily)	σ_j^D (daily)
Part1	0	1.22	NA	NA	NA	NA	28	11.22	x	x
Part2	1	0.20	NA	NA	NA	NA	15	0	x	x
Part3	2	1.36	NA	NA	NA	NA	10	0	x	x
Manu1	3	x	1.25	146	3.15	134.99	x	x	x	x
Manu2	4	x	0.19	27	6.14	35.56	x	x	x	x
Ret1	5	x	NA	NA	NA	NA	0	0	253	36.62
Ret2	6	x	NA	NA	NA	NA	0	0	45	1
Ret3	7	x	NA	NA	NA	NA	0	0	75	2

4.3.1 Impact of demand issues

We assume that the internal demand level of node Manu1 varies under different parameter settings (e.g. as the result of a variation in the external demand of node Ret1). For every parameter setting, we assume that the assembly supply chain fulfills the external demand and show the performance of manufacturing node Manu1 under these parameter settings. Studying this node is particularly interesting as its performance is both subject to production capacity constraints as well as its demand volume. Table 3 shows the queuing results of the first phase for node Manu1 under the different parameter settings. Remark that the utilization rate, the average lead time and the lead time variability increase as the node's load increases.

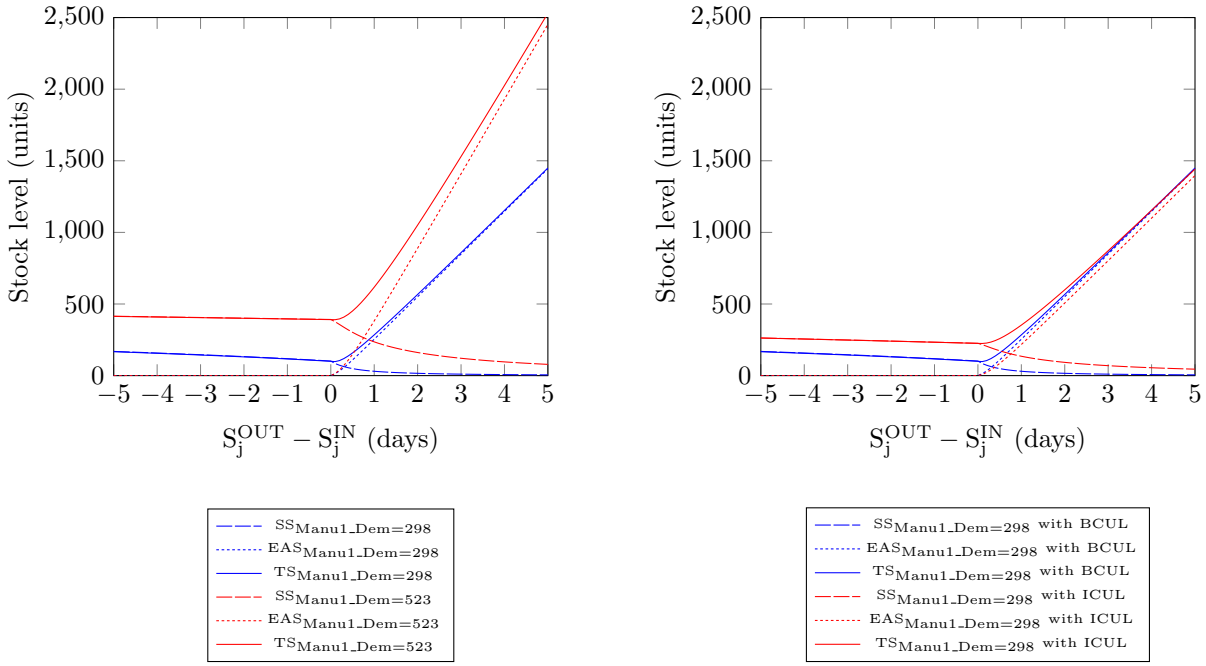
Figure 5a plots the SS, the accumulated EAS and their sum, the total stock, for a range of pairs of inbound and outbound guaranteed service times for parameter settings 1 and 10 for node Manu1. This figure illustrates the following:

- Similarly to [Humair et al., 2013], we observe for both parameter settings that the necessary SS to guarantee the quoted outbound service time increases as the difference between the outbound service time and inbound service time becomes more negative and EAS piles up as the difference

Table 3: Results for node Manu1 ($j = 3$) under different parameter settings

Parameter setting	Parameter setting number	ρ_j	ew_j^Q (days)	σ_j^{ewQ} (daily)	ew_j (days)	σ_j^{ew} (daily)
Manu1_Dem = 298	1	0.5215	0.1221	0.4677	0.2619	0.4695
Manu1_Dem = 323	2	0.5653	0.1328	0.4780	0.2715	0.4797
Manu1_Dem = 348	3	0.6090	0.1448	0.4896	0.2827	0.4912
Manu1_Dem = 373	4	0.6527	0.1585	0.5027	0.2955	0.5042
Manu1_Dem = 398	5	0.6965	0.1742	0.5174	0.3105	0.5189
Manu1_Dem = 423	6	0.7402	0.1924	0.5341	0.3279	0.5355
Manu1_Dem = 448	7	0.7840	0.2137	0.5532	0.3486	0.5545
Manu_Dem = 473	8	0.8278	0.2390	0.5751	0.3732	0.5764
Manu1_Dem = 498	9	0.8715	0.2696	0.6006	0.4031	0.6018
Manu1_Dem = 165	10	0.8129	0.3070	0.6307	0.4400	0.6318

Figure 5: Impact of demand and capacity issues for node Manu1



(a) Impact of demand issues

(b) Impact of capacity issues

between the outbound and inbound guaranteed service times becomes more positive. This leads to the nonconcavity of the node's total stock ([Humair et al., 2013]).

- When we compare the two parameter settings, we observe that the node's total stock curve of parameter setting 10 (marked in red) lies on top of parameter setting 1 (marked in blue). As the node's load increases, the impact on the supply chain's the total supply chain stock is negative.

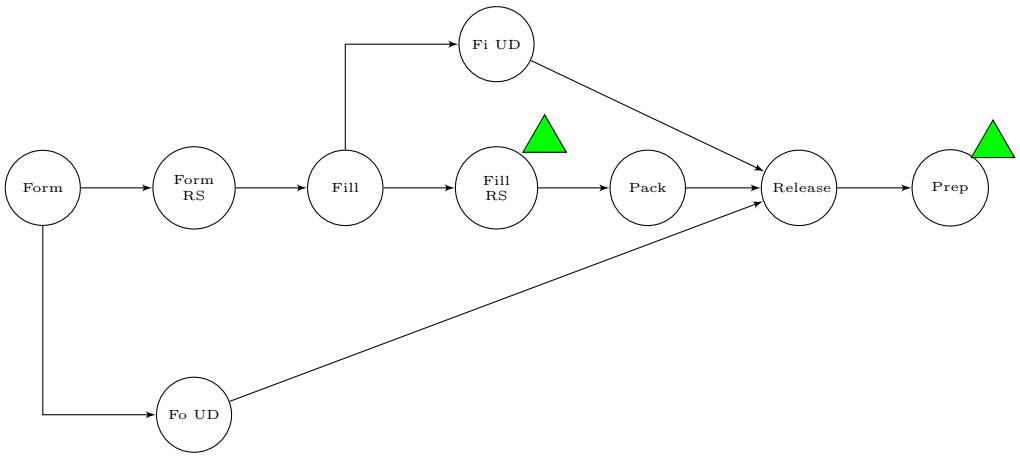
4.3.2 Impact of capacity issues

In Figure 5b, we show the SS, EAS and total stock for parameter setting 1 with a base level and an increased level of the production capacity utilization. We are interested in the impact of the production capacity while these stock levels are independent of the change of the node's internal demand. Such increase in the production capacity utilization can be due to e.g. breakdowns or capacity allocated to other products. Figure 5b is constructed as follows:

- The plots for the Base Capacity Utilization Level (BCUL) are the same to those in Figure 5a (marked in blue).
- For the plots with an Increased Capacity Utilization Level (ICUL), we use the same input data as in parameter setting 1 except for the capacity utilization level. We use the capacity utilization level of parameter setting 10 and recompute the lead time and its variability for parameter setting 1.

The SS increases when an increased production capacity utilization level is considered because a longer and more variable lead time needs to be buffered. On the contrary, the EAS decreases for an increased production capacity utilization level. This is because, on average, less lead time realizations are smaller than the difference between the node's outbound and inbound service times as the node's average lead time and lead time variability increase. Remark that in industry both effects (demand and capacity issues) may come into play simultaneously and even further increase the importance of trading off operational buffers ([Vandaele and De Boeck, 2003]).

Figure 6: Structure of the secondary processes of the rotavirus vaccine supply chain of GSK Vaccines



5 Application to the design of a rotavirus vaccine supply chain

We are now ready to illustrate our approach for the design of a rotavirus vaccine supply chain. In [Lemmens et al., 2016] we discuss the key issues to design a vaccine supply chain and in [Decouttere et al., 2015] we motivate the choice of the rotavirus vaccine as an illustrative case. Rotaviruses are double-stranded RNA viruses that replicate in the intestine and represent the worldwide major cause of severe diarrhoea for children younger than 5 years ([Smith et al., 1980]). [Tate et al., 2012] estimate that this virus caused the deaths of 453000 children worldwide in 2008. Since 2006, two effective rotavirus vaccines have been licensed and are recommended for use in all countries by the WHO, particularly in those countries with a high diarrhoea mortality rate for children younger than 5 years.

5.1 Supply chain description

[Pujar et al., 2014] explain that the vaccine antigen is generated via a cultivation process (we refer to it as the primary process) and is characterized by an appropriate choice of cell substrate, growth media and fermentation or cell culture conditions that reproducibly produce the antigen in large quantities. Subsequently, the vaccine purification process may remove host cell impurities, process additives and yields a bulk vaccine. The bulk vaccine is converted into a final vaccine product in the formulation (Form), filling (Fill) and packaging (Pack) processes. We refer to these processes as (a subset of) the secondary processes which are shown in Figure 6. Through these stages, the vaccine is formulated into a final composition which can be in liquid or lyophilized form, and then presented in an appropriate final presentation, such as plastic tubes, vials or prefilled oral applicators.

The quality control and quality assurance department guarantees persistence of quality during these processes by various tests. Some of these tests are performed in parallel with the subsequent manufacturing stages as the lead time of these tests can be long. However, to continue the manufacturing processes, some critical tests need to be performed immediately after the formulation and filling stages to identify whether the vaccines are allowed to proceed to the next stage. When the results of these critical tests are positive, the vaccines are put in a Restricted Status (Form RS and Fill RS). To release the vaccine, the outcome of the Usage Decisions (Form UD and Fill UD) performed in parallel need to be positive. After the release stage vaccines are prepared (Prep) for shipment and wait in cold storage until they can be shipped to local warehouses worldwide.

Especially for developing countries, an appropriate presentation to administer the vaccine is assessed when introducing a vaccine in a country's immunization programme. For the rotavirus vaccine the volume per box of vaccines depends on the vaccine presentation: squeezable tubes, oral applicators or vials ([WHO, 2013]). These volumes are an important determinant to decide on the use of one of these presentations as the cold chain capacity limits the number of doses that can be transported and stored in refrigerators. However, vaccines also need to be easy to administer as not all of the health care workers and volunteers receive the same training ([WHO, 2014]). Generally, the plastic tube presentation of the rotavirus vaccine is perceived as an easy to use presentation as it can be administered in a swift and safe way. This presentation has been introduced in many developing countries and implies that the majority

Table 4: Lead time data and guaranteed service time solution for the base case

Stage name	ew_j (days)	σ_j^{ew} (daily)	GST_j^{IN} (days)	GST_j^{OUT} (days)
Form	1	0	100	101
Form RS	1	0	101	102
Fill*	3.04	1.74	102	105.04
Fill RS	59.03	28.91	105.04	213.02
Pack*	0.15	0.40	213.02	213.17
Release	23.47	11.21	213.17	236.64
Prep	81.67	36.31	236.64	280
Fo UD	112.17	31.9	101	213.17
Fi UD	99.18	50	105.04	204.22

of the demand volume are tubes. Therefore we focus on the tube presentation of the vaccine in the remainder of this work.

5.2 Numerical illustration

We use inequality (1) to model the quality tests performed in parallel with the manufacturing stages. We model the release of a batch of vaccines as an assembly process of the manufacturing stages and the outcome of the Usage Decisions as the release is only allowed to proceed in case of the last of the three preceding nodes has finished successfully. Remark that inequality (1) is also relevant for combination vaccines (e.g. DTP vaccine, [WHO, 2016]): all the vaccine components need to be ready before the manufacturing process can continue.

For our numerical illustration, we focus on the secondary processes and assume that the primary process takes 100 days. For the secondary processes we extracted the lead times for a sample of 103 batches using GSK Vaccines' SAP system. Table 4 shows the stages' average lead time and corresponding variability. Remark that an asterisk indicates a supply chain stage for which the lead time and its variability are computed in Phase 1 using the proposed queuing network methodology. In this table, we distinguish the processes performed in parallel with the manufacturing stages by a dashed line.

In addition, we were provided with the following production capacity data. A formulated batch consists of 270000 units to be processed. The filling and packaging stages contain high-speed batch processing activities and consist of three and two parallel lines respectively. These lines are configured in a way such that each filling line processes 200 units per minute and each packaging line processes 329 units per minute. For the filling lines as well as the packaging lines, we assume that they operate 3 shifts of 8 hours per day and 5 days per week. On average, 64 % of the daily operating time is dedicated to actual processing whilst the remaining percentage is used as setup time (e.g. maintenance, validation). Furthermore we assume that GSK Vaccines processes 240 formulated batches in a year and the yearly demand volume to be normally distributed with a mean of 60 million doses and a standard deviation of 61785 doses.

The introduction of the rotavirus vaccine in a new country may lead to a substantial increase of the demand volume. In this paper, we are interested in the case that the demand volume may increase from 60 mio doses (base case) to 65 mio doses and that the annual number of manufactured batches increases from 240 batches to 260 batches accordingly. For this illustrative case, we see these values as the possible realizations for the demand volume as an event parameter.

Furthermore, GSK Vaccines is interested in cutting lead times as the company is setting up lead time reduction programmes to gradually reduce its lead times. We will demonstrate the impact of cutting lead times on the supply chain stock and compare the results with the current lead times. We assume that GSK Vaccines has the option to shorten (1) Fo UD with 30 days or (2) Fi UD with 30 days or (3) Fill RS with 30 days. We see these values as the possible realizations for the lead times as a design parameter.

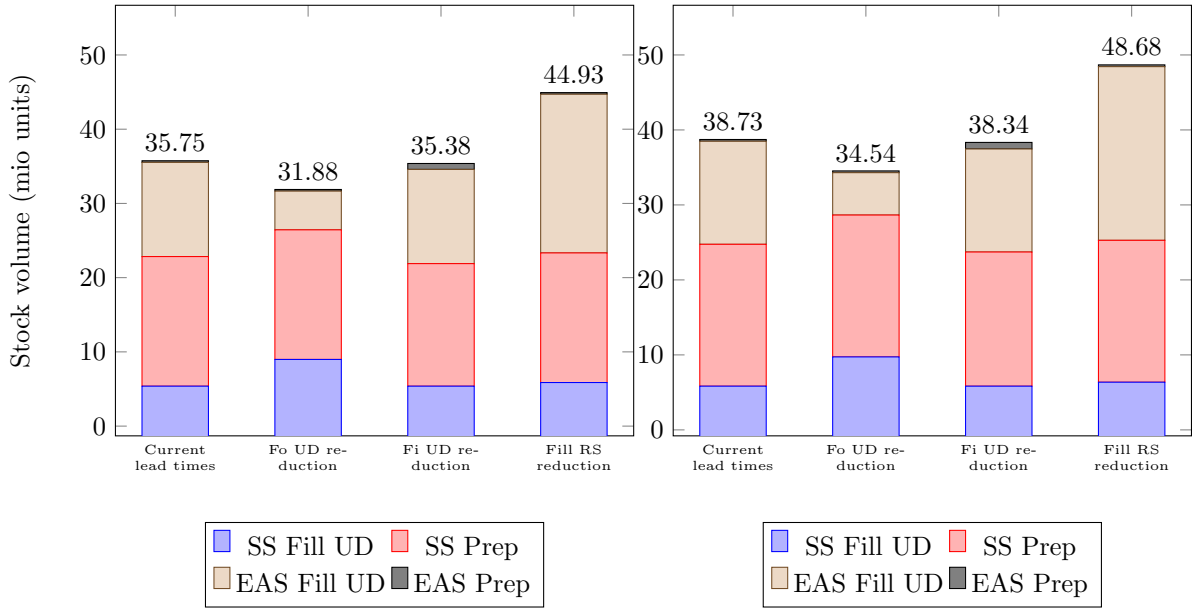
Table 5 shows the results for the production capacity utilization for filling and packaging. For completeness we show the results for 8 scenarios for our model-KPI from Phase 1. Note that this number of scenarios is the result of a full factorial design of the possible realizations for our event (2) and design parameter (4).

We assume that the supply chain responsiveness is a KPI which comes straight from the scenario description and equals 280 days. For the base scenario (current lead times and demand volume) we show the guaranteed service time solution in Table 4. For each scenario, Figure 7 shows the results of the inventory optimization of Phase 2. Panel 7a shows that the largest benefits of cutting lead times

Table 5: Results for the production capacity utilization of Phase 1

		Event parameter			
		Current demand volume		Increased demand volume	
		Filling	Packaging	Filling	Packaging
Design parameter	Current lead times	63.18 %	72.72 %	68.45 %	78.78 %
	Fo UD reduction	63.18 %	72.72 %	68.45 %	78.78 %
	Fi UD reduction	63.18 %	72.72 %	68.45 %	78.78 %
	Fill RS reduction	63.18 %	72.72 %	68.45 %	78.78 %

Figure 7: Impact of reducing lead times at Fo UD, Fi UD and Fill RS with 30 days



(a) Current demand volume (60 mio doses)

(b) Increased demand volume (65 mio doses)

can be obtained in case of reducing the Fo UD stage with 30 days: the yearly stock decreases with 3.87 mio doses (10.83 %). This is because the reduction of Fo UD has the largest impact on the decrease of the inbound guaranteed service time of the release stage and subsequently leads to the largest reduction in EAS before the release stage. For the scenario representing a reduction of Fill RS with 30 days, we observe an increase of the EAS and the supply chain stock. This is due to the fact that the release process is seen as assembly stage and its inbound guaranteed service time does not change in this scenario. The lead time reduction of Fill RS leads to especially more EAS before the release stage. For Panel 7b we observe that the total stock for each scenario is larger compared to its corresponding scenario in Panel 7a. Note that this effect is partly due to the integration of the impact of the production capacity on the lead time (Phase 1). Also for this panel we observe that the largest benefits of cutting lead times can be obtained in case of reducing the Fo UD stage with 30 days: the yearly stock decreases with 4.19 mio doses (12.13 %).

For this illustrative case we observe that the design parameter realization "cutting lead times of Fo UD with 30 days" shows promising results according to our KPI in case of both event parameter realizations. Note that the extraction of a design parameter realization which shows the best results in case of all events can become more complicated as the number of design parameters and event parameters increases.

6 Conclusions and further steps

This work extends the guaranteed service approach with variable lead times by the integration of production capacity. Therefore we decompose a manufacturing node into a batch processing activity and compute the lead time and its variability for such a node by the use of G/G/m queuing networks. This

extension allows a supply chain design to be evaluated according to multiple performance measures: supply chain inventories, supply chain responsiveness and production capacity utilization.

We extended 12 instances of the data set of [Willems, 2008] to validate our approach. For our extended data set, all the manufacturing stages are modeled with queuing networks to calculate their lead time and lead time variability. Similarly to [Humair et al., 2013] we observe that the number of SS locations increases as the number of stages with variable lead times increases to avoid EAS ramifications.

We see three opportunities to extend our modeling approach. A first step of this research is to decompose a manufacturing node into a (sub)network of nodes. Such an approach allows us to measure the impact of different batch sizes (under regulatory and manufacturing constrains) on the lead time ([Lambrecht and Vandaele, 1996, Lambrecht et al., 1998]). A second opportunity is to model the lead times of the quality processes as endogenous variables as these processes are time-consuming. Finally, another extension is to integrate shared capacity between different vaccines. This extension is highly relevant for more complex manufacturing vaccine supply chains as production capacity is carefully allocated between multiple products to satisfy the demand of a set of vaccines.

Figure 1 shows that our scenarios will correspond to a combination of design parameters and event parameters for the vaccine supply chain. We are interested in finding design parameters that are sufficiently robust with respect to uncertain future events. Therefore we will evaluate the performance of a set of scenarios according to two types of KPIs: model-based and non-model KPIs ([Decouttere et al., 2015, Hahn et al., 2016]). Figure 4 shows the two phases of our flow model and the positioning of this paper's contribution to reveal both undesirable vis-a-vis championing design parameters by evaluating the model-based KPIs of each scenario. To satisfy the needs of the vaccine manufacturer and its stakeholders for such a design parameter, an appropriate reconciliation of economical, technological and value-based KPIs is required ([Decouttere et al., 2015]). We refer the interested reader into the performance evaluation of set of scenarios according to model-based and non-model KPIs and the determination of a scenario's efficiency and scenario ranking to [Decouttere et al., 2015] and [Hahn et al., 2016].

We emphasized the importance of studying the lead times for the design of a vaccine supply chain. Both literature and industry characterize the vaccine supply chain as unresponsive and show their interest to reduce lead times. During past few years, GSK Vaccines has shifted more attention from demand forecasting toward lead time reduction: the company has implemented various initiatives to map the production lead times and to identify the critical chain of its vaccines. Furthermore, GSK Vaccines is currently considering to implement a company wide lead time reduction programme. We considered such lead time reduction programmes as a design parameter being part of the scenarios of our numerical illustration. We demonstrated the extraction of a design parameter realization in case of the current demand volume as well as an increased demand volume. We note that such lead time reductions may not be straightforwardly achievable and take several years to implement as process changes require a stringent validation by the relevant regulatory authorities.

For our industrial application, we will build more complex scenarios and evaluate them according to multiple KPIs. A more complex design parameter may, for example, encompass the production of more thermostable vaccines ([Chen and Kristensen, 2009]). In [Decouttere et al., 2015] we show how to validate complex scenarios with industry experts as well as with relevant stakeholders. Finally, we point out that our generic approach can be extended to other, more complex vaccine manufacturing supply chains as well as to other health supply chains and even to supply chain design in general.

Acknowledgment

We gratefully acknowledge financial support from the GlaxoSmithKline Vaccines Research Chair on Operations Management and Re-Design of Healthcare Supply Chains in Developing Countries to increase Access to Medicines. We also thank Brian T. Tomlin and Sean P. Willems for their useful comments on some technical aspects of their paper ([Humair et al., 2013]).

Appendix A

We show the equations to calculate the SS and EAS for GSA-VAR. The equations exposed in this appendix can be considered as complementary to the ones described in [Humair et al., 2013].

$$SS_j = k_j \sqrt{Y(T)(\sigma_j^D)^2 + (\mu_j^D)^2 Z(T)} \tag{37}$$

$$EAS_j = \mu_j^D (Y(T) - ew_j + T) \tag{38}$$

where $T = S_j^{OUT} - S_j^{IN}$ and functions Y() and Z() are specified as follows:

$$Y(T) = H_j^2(T^+) - T(1 - H_j^1(T^+)) \tag{39}$$

$$Z(T) = T^2 H_j^1(T^+)(1 - H_j^1(T^+)) - 2T H_j^1(T^+) H_j^2(T^+) + H_j^3(T^+) - (H_j^2(T^+))^2 \tag{40}$$

where $T^+ = [S_j^{OUT} - S_j^{IN}]^+ = \max\{S_j^{OUT} - S_j^{IN}, 0\}$. In case of a discrete lead time distribution for stage j, the H-functions are specified as follows:

$$H_j^1(T^+) = \sum_{l=0}^{T^+} \pi[L_j = l] \tag{41}$$

$$H_j^2(T^+) = \sum_{l=T^+}^{\infty} l\pi[L_j = l] \tag{42}$$

$$H_j^3(T^+) = \sum_{l=T^+}^{\infty} l^2\pi[L_j = l] \tag{43}$$

where $\pi[L_j = l]$ denotes the probability that the realized lead time for stage j equals l. For a continuous lead time distribution, the H-functions become $H_j^1(T^+) = \int_{l=0}^{T^+} f_{L_j}(l)dl$, $H_j^2(T^+) = \int_{l=T^+}^{\infty} l f_{L_j}(l_j)dl$ and $H_j^3(T^+) = \int_{l=T^+}^{\infty} l^2 f_{L_j}(l)dl$ with $f_{L_j}(l)$ the lead time distribution function.

Appendix B

We use the approximation of [Whitt, 1993] to calculate the variance of the batch waiting time:

$$(\sigma_j^{ew^Q})^2 = (ew_j^Q)^2 (C_j^{ew^Q})^2 \tag{44}$$

$$(C_j^{ew^Q})^2 = ((C_j^W)^2 + 1 - \pi_j^D)/\pi_j^D \tag{45}$$

$$\pi_j^D = \rho_j + ((C_j^{BA})^2 - 1)\rho_j(1 - \rho_j)h_j(\rho_j, (C_j^{BA})^2, (C_j^{BS})^2) \tag{46}$$

$$h_j(\rho_j, (C_j^{BA})^2), (C_j^{BS})^2) = \begin{cases} \frac{1+(C_j^{BA})^2+\rho_j[C_e^2]_j}{1+\rho_j((C_j^{BS})^2-1)+\rho_j^2(4(C_j^{BA})^2+(C_j^{BS})^2)}, & (C_j^{BA})^2 \le 1 \\ \frac{4\rho_j}{(C_j^{BA})^2+\rho_j^2(4(C_j^{BA})^2+(C_j^{BS})^2)}, & (C_j^{BA})^2 \ge 1 \end{cases} \tag{47}$$

$$(C_j^W)^2 = 2\rho_j - 1 + \frac{4}{3}(1 - \rho_j)d_j^3/((C_j^{BS})^2 + 1)^2 \tag{48}$$

$$d_j^3 = \begin{cases} \frac{3}{4}[\frac{1}{q_j^2} + \frac{1}{(1-q_j)^2}], & (C_j^{BS})^2 \ge 1 \\ (2(C_j^{BS})^2 + 1)((C_j^{BS})^2 + 1), & (C_j^{BS})^2 < 1 \end{cases} \tag{49}$$

$$q_j = [1 + \sqrt{4((C_j^{BS})^2 - 1)((C_j^{BS})^2 + 1)}]/2 \tag{50}$$

where $(C_j^{ew^Q})^2$ is the SCV of the batch waiting time, π_j^D is the probability of delay or the probability that an arrival must wait before service can start and $(C_j^W)^2$ is the SCV of the batch waiting time given that the machine is busy.

References

[ATM, 2016] ATM (2016). Access to medicine index. http://www.accesstomedicineindex.org/. Accessed: 2016-03-17.

[Billington et al., 2004] Billington, C., Callioni, G., Crane, B., Ruark, J. D., Rapp, J. U., White, T., and Willems, S. P. (2004). Accelerating the profitability of hewlett-packard's supply chains. *Interfaces*, 34(1):59–72.

[Bossert and Willems, 2007] Bossert, J. and Willems, S. P. (2007). A periodic-review modeling approach for guaranteed service supply chains. *Interfaces*, 37(5):420–435.

[Chen and Kristensen, 2009] Chen, D. and Kristensen, D. (2009). Opportunities and challenges of developing thermostable vaccines. *Expert Review of Vaccines*, 8(5):547–557.

[Chen and Li, 2015] Chen, H. and Li, P. (2015). Optimization of (r, q) policies for serial inventory systems using the guaranteed service approach. *Computers & Industrial Engineering*, 80(0):261–273.

[Croxton et al., 2003] Croxton, K. L., Gendron, B., and Magnanti, T. L. (2003). A comparison of mixed-integer programming models for nonconvex piecewise linear cost minimization problems. *Management Science*, 49(9):1268–1273.

[de Treville et al., 2014] de Treville, S., Bicer, I., Chavez-Demoulin, V., Hagspiel, V., Schürhoff, N., Tasserit, C., and Wager, S. (2014). Valuing lead time. *Journal of Operations Management*, 32(6):337–346.

[Decouttere et al., 2015] Decouttere, C. J., Vandaele, N. J., Lemmens, S., and Bernuzzi, M. (2015). *Advances in Managing Humanitarian Operations*, chapter The Vaccine Supply Chain Multathlon: the Reconciliation of Technology, Economy and Access to Medicines, pages 205–227. Springer International Series in Operations Research & Management Science.

[Eruguz et al., 2014] Eruguz, A. S., Jemai, Z., Sahin, E., and Dallery, Y. (2014). Optimising reorder intervals and order-up-to levels in guaranteed service supply chains. *International Journal of Production Research*, 52(1):149–164.

[Eruguz et al., 2016] Eruguz, A. S., Sahin, E., Jemai, Z., and Dallery, Y. (2016). A comprehensive survey of guaranteed-service models for multi-echelon inventory optimization. *International Journal of Production Economics*, 172:110–125.

[Farasyn et al., 2011] Farasyn, I., Humair, S., Kahn, J. I., Neale, J. J., Rosen, O., Ruark, J., Tarlton, W., de Velde, W. V., Wegryn, G., and Willems, S. P. (2011). Inventory optimization at procter & gamble: Achieving real benefits through user adoption of inventory tools. *Interfaces*, 41(1):66–78.

[Funaki, 2012] Funaki, K. (2012). Strategic safety stock placement in supply chain design with due-date based demand. *International Journal of Production Economics*, 135(1):4–13.

[Grahl et al., 2014] Grahl, J., Minner, S., and Dittmar, D. (2014). Meta-heuristics for placing strategic safety stock in multi-echelon inventory with differentiated service times. *Annals of Operations Research*, pages 1–16.

[Graves and Schoenmeyr, 2016] Graves, S. C. and Schoenmeyr, T. (2016). Strategic safety-stock placement in supply chains with capacity constraints. *Manufacturing & Service Operations Management*, 18(3):445–460.

[Graves and Willems, 2000] Graves, S. C. and Willems, S. P. (2000). Optimizing strategic safety stock placement in supply chains. *Manufacturing & Service Operations Management*, 2(1):68–83.

[Graves and Willems, 2003] Graves, S. C. and Willems, S. P. (2003). Supply chain design: Safety stock placement and supply chain configuration. In Graves, S. C. and de Kok, A. G., editors, *Supply Chain Management: Design, Coordination and Operation*, volume 11 of *Handbooks in Operations Research and Management Science*, pages 95–132. Elsevier.

[Graves and Willems, 2005] Graves, S. C. and Willems, S. P. (2005). Optimizing the supply chain configuration for new products. *Management Science*, 51(8):1165–1180.

[Graves and Willems, 2008] Graves, S. C. and Willems, S. P. (2008). Strategic inventory placement in supply chains: Nonstationary demand. *Manufacturing & Service Operations Management*, 10(2):278–287.

[GSK, 2014] GSK (2014). Gsk and the decade of vaccines. http://www.gsk.com/media/281058/gsk-and-the-decade-of-vaccines-report.pdf. Accessed: 2016-03-17.

[GSK, 2016a] GSK (2016a). Gsk expands graduated approach to patents and intellectual property to widen access to medicines in the worlds poorest countries. http://www.gsk.com/en-gb/media/press-releases/2016/gsk-expands-graduated-approach-to-patents-and-intellectual-property-to-widen-access-to-medicines-in-the-world-s-poorest-countries/. Accessed: 2016-05-27.

[GSK, 2016b] GSK (2016b). Vaccines. http://www.gsk.com/en-gb/about-us/what-we-do/vaccines/. Accessed: 2016-03-17.

[Hahn et al., 2016] Hahn, G. J., Sens, T., Decouttere, C. J., and Vandaele, N. J. (2016). A multi-criteria approach to robust outsourcing decision-making in stochastic manufacturing systems. *Computers & Industrial Engineering*, 98:275–288.

[Humair et al., 2013] Humair, S., Ruark, J., Tomlin, B., and Willems, S. P. (2013). Incorporating stochastic lead times into the guaranteed service model of safety stock optimization. *Interfaces*, 43(5):421–434.

[Humair and Willems, 2006] Humair, S. and Willems, S. P. (2006). Optimizing strategic safety stock placement in supply chains with clusters of commonality. *Operations Research*, 54(4):725–742.

[Humair and Willems, 2011] Humair, S. and Willems, S. P. (2011). Optimizing strategic safety stock placement in general acyclic networks. *Operations Research*, 59(3):781–787.

[IFPMA, 2016] IFPMA (2016). The complex journey of a vaccine. the manufacturing chain, regulatory requirements and vaccine availability. http://www.ifpma.org/wp-content/uploads/2016/01/TheComplexJourneyofaVaccinePRINT-EN.pdf. Accessed: 2016-03-17.

[Jung et al., 2008] Jung, J. Y., Blau, G., Pekny, J. F., Reklaitis, G. V., and Eversdyk, D. (2008). Integrated safety stock management for multi-stage supply chains under production capacity constraints. *Computers & Chemical Engineering*, 32(11):2570–2581.

[Klosterhalfen et al., 2013] Klosterhalfen, S. T., Dittmar, D., and Minner, S. (2013). An integrated guaranteed- and stochastic-service approach to inventory optimization in supply chains. *European Journal of Operational Research*, 231(1):109–119.

[Klosterhalfen and Minner, 2010] Klosterhalfen, S. T. and Minner, S. (2010). Safety stock optimisation in distribution systems: a comparison of two competing approaches. *International Journal of Logistics Research and Applications*, 13(2):99–120.

[Klosterhalfen et al., 2014] Klosterhalfen, S. T., Minner, S., and Willems, S. P. (2014). Strategic safety stock placement in supply networks with static dual supply. *Manufacturing & Service Operations Management*, 16(2):204–219.

[Kraemer and Langenbach-Belz, 1976] Kraemer, W. and Langenbach-Belz, M. (1976). Approximate formulae for the dealy in the queuing system gi/g/1. In *Congressbook, Eighth International Teletraffic Congress*, pages 235–1/8, Melbourne.

[Lambrecht et al., 1998] Lambrecht, M. R., Ivens, P. L., and Vandaele, N. J. (1998). Aclips: A capacity and lead time integrated procedure for scheduling. *Management Science*, 44(11):1548–1561.

[Lambrecht and Vandaele, 1996] Lambrecht, M. R. and Vandaele, N. J. (1996). A general approximation for the single product lot sizing model with queueing delays. *European Journal of Operational Research*, 95(1):73–88.

[Lemmens et al., 2016] Lemmens, S., Decouttere, C. J., Vandaele, N. J., and Bernuzzi, M. (2016). A review of integrated supply chain network design models: key issues for vaccine supply chains. *Chemical Engineering Research and Design*, 109C:366–384.

[Lesnaia, 2004] Lesnaia, E. (2004). *Optimizing Safety Stock Placement in General Network Supply Chains*. PhD thesis, Massachusetts Institute of Technology.

[Li and Jiang, 2012] Li, H. and Jiang, D. (2012). New model and heuristics for safety stock placement in general acyclic supply chain networks. *Computers & Operations Research*, 39(7):1333–1344.

[Li and Womer, 2008] Li, H. and Womer, K. (2008). Modeling the supply chain configuration problem with resource constraints. *International Journal of Project Management*, 26(6):646–654.

[Li et al., 2013] Li, P., Chen, H., and Che, A. (2013). Optimal batch ordering policies for assembly systems with guaranteed service. *International Journal of Production Research*, 51(20):6275–6293.

[Little, 1961] Little, J. D. C. (1961). A proof for the queuing formula: L= λ w. *Operations Research*, 9(3):383–387.

[Magnanti et al., 2006] Magnanti, T. L., Shen, Z.-J. M., Shu, J., Simchi-Levi, D., and Teo, C.-P. (2006). Inventory placement in acyclic supply chain networks. *Operations Research Letters*, 34(2):228–238.

[Moncayo-Martínez et al., 2014] Moncayo-Martínez, L. A., Reséndiz-Flores, E. O., Mercado, D., and Sánchez-Ramírez, C. (2014). Placing safety stock in logistic neworks under guaranteed-service time inventory models: An application to the autmotive industry. *Journal of Applied Research and Technology*, 12(3):538–550.

[Moncayo-Martínez and Zhang, 2013] Moncayo-Martínez, L. A. and Zhang, D. Z. (2013). Optimising safety stock placement and lead time in an assembly supply chain using bi-objective max–min ant system. *International Journal of Production Economics*, 145(1):18–28.

[Neale and Willems, 2009] Neale, J. J. and Willems, S. P. (2009). Managing inventory in supply chains with nonstationary demand. *Interfaces*, 39(5):388–399.

[Pujar et al., 2014] Pujar, N. S., Sagar, S. L., and Lee, A. L. (2014). *History of Vaccine Process Development*, pages 1–24. John Wiley & Sons, Inc.

[Schoenmeyr and Graves, 2009] Schoenmeyr, T. and Graves, S. C. (2009). Strategic safety stocks in supply chains with evolving forecasts. *Manufacturing & Service Operations Management*, 11(4):657–673.

[Shah, 2004] Shah, N. (2004). Pharmaceutical supply chains: key issues and strategies for optimisation. *Computers & Chemical Engineering*, 28(6-7):929–941.

[Shah, 2005] Shah, N. (2005). Process industry supply chains: Advances and challenges. *Computers & Chemical Engineering*, 29(6):1225–1236.

[Shu and Karimi, 2009] Shu, J. and Karimi, I. (2009). Efficient heuristics for inventory placement in acyclic networks. *Computers & Operations Research*, 36(11):2899–2904.

[Sitompul et al., 2008] Sitompul, C., Aghezzaf, E.-H., Dullaert, W., and Landeghem, H. V. (2008). Safety stock placement problem in capacitated supply chains. *International Journal of Production Research*, 46(17):4709–4727.

[Smith et al., 1980] Smith, M. L., Lazdins, I., and Holmes, I. H. (1980). Coding assignments of double-stranded rna segments of sa 11 rotavirus established by in vitro translation. *Journal of Virology*, 33(3):976–982.

[Suri, 1998] Suri, R. (1998). *Quick response manufacturing : A companywide approach to reducing lead times.* Chicago: American technical society.

[Tate et al., 2012] Tate, J. E., Burton, A. H., Boschi-Pinto, C., Steele, A. D., Duque, J., and Parashar, U. D. (2012). 2008 estimate of worldwide rotavirus-associated mortality in children younger than 5 years before the introduction of universal rotavirus vaccination programmes: a systematic review and meta-analysis. *The Lancet Infectious Diseases*, 12(2):136–141.

[VaccinesEurope, 2015] VaccinesEurope (2015). How are vaccines produced? http://www.vaccineseurope.eu/about-vaccines/key-facts-on-vaccines/how-are-vaccines-produced/. Accessed: 2016-03-17.

[Vandaele, 1996] Vandaele, N. J. (1996). *The Impact of Lot Sizing on Queueing Delays: Multi Product, Multi Machine Models.* PhD thesis, KU Leuven.

[Vandaele and De Boeck, 2003] Vandaele, N. J. and De Boeck, L. (2003). Advanced resource planning. *Robotics and Computer-Integrated Manufacturing*, 19(1-2):211–218.

[Whitt, 1983] Whitt, W. (1983). The queueing network analyzer. *The Bell System Technical Journal*, 62(9):2779–2815.

[Whitt, 1993] Whitt, W. (1993). Approximations for the gi/g/m queue. *Production and Operations Management*, 2(2):114–161.

[WHO, 2013] WHO (2013). Information of rotavirus vaccines: Information for policy makers, programme managers, and health workers. http://www.who.int/immunization/monitoring_surveillance/burden/vpd/surveillance_type/sentinel/rotavirus_intro_guidance_who_july31_2013.pdf. Accessed: 2016-03-17.

[WHO, 2014] WHO (2014). Principles and considerations for adding a vaccine to a national immunization programme: From decision to implementation and monitoring. http://apps.who.int/iris/bitstream/10665/111548/1/9789241506892_eng.pdf. Accessed: 2016-03-17.

[WHO, 2016] WHO (2016). Pertussis - the vaccine. http://www.who.int/immunization /topics/pertussis/en/index2.html. Accessed: 2016-05-09.

[Wieland et al., 2012] Wieland, B., Mastrantonio, P., Willems, S. P., and Kempf, K. G. (2012). Optimizing inventory levels within intel's channel supply demand operations. *Interfaces*, 42(6):517–527.

[Willems, 2008] Willems, S. P. (2008). Data set—real-world multiechelon supply chains used for inventory optimization. *Manufacturing & Service Operations Management*, 10(1):19–23.

[You and Grossmann, 2010] You, F. and Grossmann, I. E. (2010). Integrated multi-echelon supply chain design with inventories under uncertainty: Minlp models, computational strategies. *American Institute of Chemical Engineers Journal*, 56(2):419–440.

[You and Grossmann, 2011a] You, F. and Grossmann, I. E. (2011a). Balancing responsiveness and economics in process supply chain design with multi-echelon stochastic inventory. *American Institute of Chemical Engineers Journal*, 57(1):178–192.

[You and Grossmann, 2011b] You, F. and Grossmann, I. E. (2011b). Stochastic inventory management for tactical process planning under uncertainties: Minlp models and algorithms. *American Institute of Chemical Engineers Journal*, 57(5):1250–1277.

FACULTY OF ECONOMICS AND BUSINESS
Naamsestraat 69 bus 3500
3000 LEUVEN, BELGIË
tel. + 32 16 32 66 12
fax + 32 16 32 67 91
info@econ.kuleuven.be
www.econ.kuleuven.be